Research Note

POLLEX-Online: The Polynesian Lexicon Project Online

Simon J. Greenhill and Ross Clark

UNIVERSITY OF AUCKLAND

The Polynesian lexicon project, POLLEX, was initiated in 1965 by Bruce Biggs in order to provide a large-scale comparative dictionary of Polynesian languages. Since then, POLLEX has grown to include over 55,000 reflexes of more than 4,700 reconstructed forms in 68 languages. These data have enabled many fundamental advances in Polynesian linguistics and prehistory. At almost half a century old, POLLEX is one of the longest-standing databases of linguistic information, and has moved through various incarnations, from typewriter and edge-punched cards, through microfiche to mainframe computer. In the last few years, online databases of linguistic information have become increasingly more prevalent, representing a major shift in the way linguistics is conducted. Online databases provide many advantages over the older forms of data distribution, including high availability, more robust data storage, and easy data manipulation and searching, and they also facilitate the replication of previous studies. This paper announces the latest reincarnation of the POLLEX database as an online resource, POLLEX-Online (http://pollex.org.nz), and describes the technical implementation details.

1. INTRODUCTION[1]

> But *whence* came the inhabitants of Polynesia? *How* did they come, or get possession of so many islands scattered over such a vast extent of ocean? *When* did they come? And *why* did they come? are questions that cannot now be answered without much conjecture. Yet, no doubt a careful and thorough examination of the several dialects, and a comparison of one with the other with a view to ascertain the groundwork of the general language, and a comparison with the languages of the neighboring continents, would not only be a subject of inquiry full of interest, but would go far to indicate the probable origin of this people. (Andrews 1836:12–13).

The Polynesians have long occupied a central position in European awareness of Oceania and its peoples: even today, scientific news that actually bears on the entire expansion of Austronesian languages into the Pacific is often presented as an answer to the perennial question of the "origin of the Polynesians."[2] The manifest lexical affinity of

1. We would like to thank Liz Pascal and Annik van Toledo for comments on the manuscript.
2. See, for example, Soares et al. (2011). This paper is reported under the headline "Genetic study uncovers new path to Polynesia," at http://www.sciencedaily.com/releases/2011/02/110203124726.htm.

Oceanic Linguistics, Volume 50, no. 2 (December 2011)
© by University of Hawai'i Press. All rights reserved.

these languages confirmed to early European visitors the unity of a people spread over much of the world's largest ocean. No wonder, then, that Lorrin Andrews expected comparative investigation to be particularly fruitful.

Lexical comparison of Polynesian languages, in fact, could be said to have begun with Hadrianus Relandus's recognition that the two brief vocabularies collected a century earlier by Jacob Le Maire at Futuna and Niuatoputapu both represented forms of the "Malay" language (Relandus 1708; see also Rensch 2000: 313–28). More substantial comparative word-lists followed in the late eighteenth century. J. R. Forster's table compared 46 words in a dozen languages, including Tahitian, Tongan, Māori, Easter Island, and Marquesan (Forster 1778). A quantitative and qualitative leap forward was made by Horatio Hale (1846), who presented about 1,000 cognate sets from ten Polynesian languages, each set headed by a "primitive or radical form" that amounts to a Proto-Polynesian reconstruction, following regular sound correspondences described in Hale's comparative grammar. Later in the century, Edward Tregear's comparative dictionary (1891) deserves mention for the large amount of material it brought together, but (in addition to being frankly Māori-centered) it made no effort to follow regular correspondences, and thus included a fair amount of spurious material.

Such was the background to the Polynesian Lexicon project (henceforth POLLEX), initiated by Bruce Biggs at the University of Auckland in 1965, using funding obtained through a New Zealand Government Lottery Grant, and a National Science Foundation Grant to the Bernice P. Bishop Museum (Pawley 2001). First fruits of the project were presented in a 1966 monograph, with D. S. Walsh as coauthor, and contributions from a large number of colleagues and students (Walsh and Biggs 1966). They presented over 900 reconstructed forms with supporting evidence from 11 Polynesian languages. While from a quantitative standpoint this might seem a small advance from what Hale had achieved, POLLEX had the advantage of both the superior data that had become available during the intervening century, and an improved reconstruction of Proto-Polynesian (PPN) that followed Elbert (1953) in recognizing PPN *h, *ʔ, and *r for the first time. The inevitable continued growth from this beginning was marked by the "interim" issue of a listing of headwords alone (Biggs, Walsh, and Waqa 1970), and a third version incorporating microfiches to handle the ever expanding corpus (Biggs 1979). By this time, POLLEX had assumed electronic form, and as paper printouts became prohibitively bulky, copies were distributed on a person-to-person basis as text files—first on floppy disks, and later as email attachments. Biggs continued to work on the project until his death in 2000, when Ross Clark assumed overall direction.

As readers of this journal well know, POLLEX has already proved to be very useful for enlightening the "intellectual darkness" surrounding Polynesian prehistory. The POLLEX database has been used to help elucidate the terms for Proto-Oceanic meteorological phenomena (Ross 1995), to understand the development of the verb 'swallow' in Oceanic languages (Lynch 2001), to reevaluate the evidence for Proto-Polynesian *h (Rutter 2001), to support new observations on the Proto-Oceanic labiovelars (Lynch 2002a), and to uncover the origins of kava (Lynch 2002b). Jeff Marck has made use of POLLEX to investigate questions ranging from whether there was an early Polynesian "Sky Father" (Marck 1996) to a full reassessment of Polynesian subgrouping (Marck

2000). However, perhaps the most intensive use was by Kirch and Green (2001), who integrated the information in POLLEX (a "monumental achievement" [p.46]) with archaeological findings to produce a rich and detailed picture of Proto-Polynesian society.

2. LANGUAGE DATABASES. POLLEX has followed ever-advancing technology—from typewriter and edge-punched cards, through microfiche and mainframe computer, to wide dispersal on personal computers. Progression to an online database is the next natural step. Recent years have seen a major growth in online databases in a range of fields (Ellis and Attwood 2001; Greenhill, Blust, and Gray 2008b). By providing vast collections in easily accessible formats, these databases are now "as important to scientific progress today as is access to a laboratory or library" (Ellis and Attwood 2001:509). Linguistics has recently begun to follow suit, with a number of large databases going online, including the Austronesian basic vocabulary database (http://language.psy.auckland.ac.nz; Greenhill, Blust, and Gray 2008a), Blust and Trussel's Austronesian comparative dictionary (http://www.trussel2.com/acd/), the world atlas of language structures (http://wals.info; Dryer and Haspelmath 2011), and the world loanword database (http://wold.livingsources.org/; Haspelmath and Tadmor 2009).

There are at least four major benefits of these online databases (Greenhill, Blust, and Gray 2008b). First, while there is a vast amount of published linguistic data, this information is often scattered between books, journal articles, manuscripts, filing cabinets, and shoeboxes. Having it in one easily accessible location reduces the reliance on happenstance library holdings or access to unpublished material. Second, much of this unpublished information is highly fragile, often stored in field notebooks, or obsolete data-storage media. This fragility is all the more concerning, as the primary sources—the languages themselves—are often heavily endangered: it is estimated that one of the world's languages goes extinct every week (Nettle and Romaine 2000). Digitizing this material and placing it online will protect it (Greenhill, Blust, and Gray 2008b). Third, online databases make easy manipulation and filtering of data possible. This capability greatly enhances the ability of researchers to test hypotheses using the data, enables the discovery of new hypotheses, and allows the robust testing of new methods (Ellis and Attwood 2001; Greenhill, Blust, and Gray 2008b; Rausher et al. 2010). Finally, online databases facilitate the citation of data. A crucial component of science is the replication and confirmation of results. Replication is all but impossible if the raw data are not accessible. This issue has recently risen to major importance in the life sciences, and many journals now require that authors use citable online data-sources, or deposit their data in such a database (for example, Rausher et al. 2010, Fairbairn 2011). Furthermore, when the original data are made available, the original publications tend to be much more useful—and get cited more often themselves (Piwowar, Day, and Fridsma 2007).

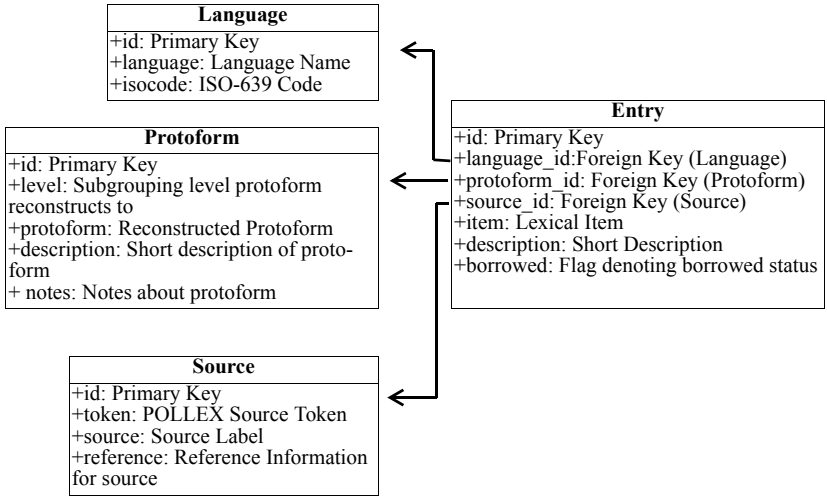
Making POLLEX available online provides these benefits: this large collection of data is in one location, is easily accessible through the internet, is safely stored in a digital format, is searchable, and is citable. This paper is thus an announcement of the availability of POLLEX-Online at http://www.pollex.org.nz. In what follows, we describe the technical implementation details of POLLEX-Online, beginning with the data layer, before moving on to describe the website and its contents.

3. TECHNICAL IMPLEMENTATION. All lexical data are stored as Unicode text (UTF-8) in the relational database system MySQL. The core of the database schema is comprised of four interlinked tables (figure 1).

1. The *Language* table stores information about the languages in Pollex. This information includes the language name, the original POLLEX three-letter identification code, and the language's ISO-639 identification code, where available.
2. The *Protoform* table stores information about POLLEX's reconstructed protoforms, the level/subgroup to which they can be reconstructed, and any notes/clarifications about the reconstructions.
3. The *Source* table contains information about the data sources incorporated into POLLEX. This includes the legacy three-letter token used to identify the source in the original POLLEX (for example, "Bge" referred to Beaglehole 1991), along with a short description of the source and a full reference.
4. The *Entry* table contains the words themselves. Each entry in this table is linked to a language in *Language* table, a protoform in the *Protoform* table, and a source in the *Source* table. Entries also have two flags: the first denoting loanwords, the second denoting problematic entries (due, for example, to phonological irregularity, dubious semantic connection, or being problematic in other ways). Each entity in these tables has all meta-information tracked, such that any changes to the data can be logged, reviewed, and reverted (undone) if necessary.

The POLLEX-Online website is implemented in the programming language *Python*, using the open-source web development framework *Django* (http://www.djangoproject.com). The user interface allows users to view the data either by language (for example, show all words from Māori), or by protoform. For example, figure 2 shows one term for the paper mulberry (*Broussonetia papyrifera*) in both the original POLLEX form (2a) and the POLLEX-Online variant (2b). The reconstruc-

FIGURE 1. SIMPLIFIED DATA SCHEME FOR POLLEX-ONLINE†



† Database tables/fields primarily for administration and logging have been excluded.

tion, Proto–Central-Eastern Polynesian *aute, has eight supporting entries in the database. These entries can be downloaded in the original POLLEX format and in a dialect of XML (suitable for computational retrieval).

FIGURE 2. ENTRIES FOR "PAPER MULBERRY (*BROUSSONETIA PAPYRIFERA*)"

a.

```
.CE    AUTE :Paper mulberry
*7     Cf. *kaute "Hibiscus".
HAW    W/auke/. :Paper mulberry
MVA    Eute. :Papyrus poussant sans culture (Jnu)
MQA    Ute. :Papyrus, murier a papier et dont on fait des etoffes (Dln)
MAO    Aute. :Bark cloth (Bgs)
PEN    Aute. :Paper mulberry (Broussonetia papyrifera) (There is none in Penrhyn) (Sta)
RAP    Aute. :Bush, used for making tapa, Broussonetia papyrifera (Sks)
RAR    `Aute. :(Broussonetia papyrifera) (Bse)
TAH    `Aute. :Broussonetia papyrifera dont on faisait un 'tapa' brun (Mte)
```

b.

Protoform: CE. AUTE - Paper mulberry

Description:	Paper mulberry
Reconstruction:	Reconstructs to CE: Central-Eastern Polynesian
Notes:	*7 Cf. *kaute "Hibiscus".

Pollex entries:

Language	Entry	Description
Hawaii	W/auke/	Paper mulberry
Mangareva	Eute	Papyrus poussant sans culture (Jnu)
Marquesas	Ute	Papyrus, murier a papier et dont on fait des etoffes (Dln)
New Zealand Maori	Aute	Bark cloth (Bgs)
Penrhyn	Aute	Paper mulberry (Broussonetia papyrifera) (There is none in Penrhyn) (Sta)
Rapa	Aute	Bush, used for making tapa, Broussonetia papyrifera (Sks)
Rarotongan	ʔAute	(Broussonetia papyrifera) (Bse)
Tahitian	ʔAute	Broussonetia papyrifera dont on faisait un 'tapa' brun (Mte)

8 entries found

Download: Pollex-Text Format, XML Format.

4. CURRENT STATISTICS. Currently there are 55,238 entries, from 68 languages and dialects, listed under 4,753 protoforms. The 68 languages include 44 varieties of Polynesian,[3] as well as collateral evidence from Polynesian's closest relatives in the Central Pacific subgroup (Eastern Fijian, Western Fijian, Rotuman), and some 21 other

Oceanic languages representing North and Central Vanuatu (6), Southeast Solomonic (10), Northwest Solomonic (2), and one each from Micronesian, Papuan Tip, and North New Guinea.

TABLE 1. LANGUAGES IN POLLEX-ONLINE, SHOWING NUMBER OF REFLEXES, IDENTIFIED LOANWORDS, AND PROBLEMATIC ENTRIES

Language	Reflexes	Loans	??	Language	Reflexes	Loans	??
’Are’are	10	0	0	Niuafo’ou	73	1	15
Aitutaki	18	0	1	Niuatoputapu	4	0	0
Anuta	772	2	65	Niuean	1,832	9	182
Arosi	41	0	8	Nuguria	701	0	57
Atiu	1	0	0	Nukumanu	108	1	6
Austral Islands	2	1	0	Nukuoro	1,349	0	84
Bugotu	54	0	3	Paamese	5	0	0
East Futunan	2,253	7	83	Penrhyn	1,520	0	142
East Uvean	1,913	9	172	Pileni	1,747	0	125
Easter Island	1,216	5	167	Pukapuka	1,880	7	319
Emae	1,003	5	58	Ra’ivavae	38	1	9
Fijian	1,169	7	242	Raga	8	0	0
Gedaged	4	0	0	Rapa	448	2	71
Gela	171	1	11	Rarotongan	2,165	5	81
Gilbertese (Kiribati)	17	0	0	Rennellese	1,807	9	136
Hawaiian	1,992	7	152	Rotuman	693	5	300
Ifira-Mele	753	4	72	Roviana	15	0	1
Kapingamarangi	1,311	1	70	Rurutu	3	0	1
Kwaio	5	0	0	Sa‘a	192	2	12
Kwara’ae	56	0	1	Samoan	2,517	5	143
Lau	110	1	13	Sikaiana	903	0	33
Lonwolwol	1	0	0	Tahitian	2,319	12	183
Luangiua	1,444	3	163	Takuu	1,065	1	39
Mangaia	232	0	18	Tikopia	1,881	4	154
Mangareva	1,318	4	88	Tokelauan	1,349	5	71
Manihiki-Rakahanga	62	0	2	Tongan	2,591	4	133
Marquesan	1,970	11	158	Toqabaqita	2	0	0
Mono-Alu	1	0	0	Tuamotuan	2,101	6	105
Moriori	866	2	64	Tupuaki	14	1	0
Mota	141	0	17	Tuvaluan	1,560	3	37
Motu	21	0	3	Vaturanga	4	0	0
Namakir	13	0	0	Wayan Fijian	249	1	27
New Zealand Māori	3,209	5	245	West Futunan	879	1	60
Ngunese	99	0	3	West Uvean	968	7	29

3. POLLEX-Online includes some data from all known Polynesian languages, including Moriori of the Chatham Islands and Niuatoputapu of northern Tonga, known only from historical documents. A few of the varieties presently distinguished in the database are better considered dialects than distinct languages. Aitutaki, Atiu, Mangaia, and Rarotongan are variants of what is commonly called Cook Islands Māori; and the Austral, Ra’ivavae, Rurutu, and Tupuaki entries probably likewise represent a single language, though the language situation in this area is much less clear. These exclusions and mergers would give a total in the mid-30s for the number of living Polynesian languages.

Table 1 shows the number of reflexes in POLLEX-Online for each language, along with the number of identified loan words and problematic entries.[4] New Zealand Māori is the best attested language with 3,209 entries (plus 5 identified loanwords and 245 problematic entries), followed by Tongan (2,591 entries, 4 loanwords, and 133 problematic entries) and Samoan (2,517 entries, 5 loanwords, and 143 problematic entries). The most widely attested protoform is *refu 'ashes' with 63 entries, followed by *futi 'pluck hair or feathers, pull up weeds, pull on a line or rope' with 51 entries, and *muri 'behind, after, to follow, be last' with 50 entries. Of the 55,238 entries in POLLEX-Online, 167 are marked as known loanwords, and 4,434 (8.01 percent) are flagged as problematic.

The reconstructions in POLLEX-Online are linked to a specific language subgroup. Table 2 shows the ten best-attested subgroups by number of reconstructed protoforms and reflexes. Unsurprisingly, the Polynesian subgroup has the most reconstructions with 1,538 identified protoforms containing 19,789 entries. The next most well-attested subgroups are Central-Eastern Polynesian, with 553 protoforms from 3,339 entries, and Nuclear Polynesian, with 469 protoforms identified from 5,022 entries.[5]

The data in POLLEX-Online are sourced from 199 different resources, ranging from dictionaries, through manuscripts, to personal communications. The most prevalent sources in POLLEX-Online are Williams (1971) with 2,691 entries, Churchward (1959) with 2,469 entries, Pratt (1911) with 2,438 entries, Lemaître (1973) with 2,163 entries, and Stimson (1964) with 2,069 entries.

5. CONCLUSION. At almost half a century old, POLLEX is one of the longest-standing databases of linguistic information. POLLEX has moved through various incarnations, from typewriter and edge-punched cards, through microfiche, to mainframe com-

TABLE 2. RECONSTRUCTED PROTOFORMS AND NUMBER OF REFLEXES FOR THE TEN BEST-ATTESTED SUBGROUPINGS IN THE POLLEX-ONLINE DATABASE

Subgroup	Reconstructions	Reflexes
Polynesian	1,538	19,789
Central-Eastern Polynesian	553	3,339
Nuclear Polynesian	469	5,022
Fijic	263	3,123
Tahitic	211	939
Oceanic	202	4,438
Austronesian	192	5,192
Malayo-Polynesian	184	4,149
Eastern Polynesian	159	1,246
Samoic-Outlier Polynesian	151	1,124

4. "Problematic" is a provisional cover term for words considered to be possible reflexes of a reconstructed form, but exhibiting unexplained semantic or phonological deviations. Some of the latter may be the result of borrowing. Further improvements to the database will include more precise annotation of these problems. In table 1, the column headed "??" consists of problematic entries.

5. A small percentage of reconstructions (perhaps eight percent of the total) have reflexes whose distribution does not clearly correspond to any established subgroup. Some of these are labeled with ad hoc distribution codes (such as XO where reflexes are found only in Outlier languages), others frankly recognized as mysteries (??).

puter. This latest incarnation places POLLEX online as a publicly available, highly accessible online database of lexical information. It is our hope that POLLEX-Online continues the strong tradition of its predecessors in helping to generate new insights into all aspects of Polynesian linguistics and prehistory. Additions and improvements to the database will continue as before. We invite readers of *Oceanic Linguistics* to sample POLLEX-Online, and provide suggestions as to how it can be made a more useful research tool.

REFERENCES

Andrews, Lorrin. 1836. Remarks on the Hawaiian dialect of the Polynesian language. *Chinese Repository* 5(2):12–21. Online at http://books.google.com/books?id=l5VCAAAAYAAJ&lpg=PA12&ots=jv-ZGJZpuo&pg=PA12

Beaglehole, Ernest, and Pearl Beaglehole. 1991. *Pukapukan dictionary manuscript*. Auckland: Pukapuka Dictionary Project, Department of Anthropology, University of Auckland.

Biggs, Bruce. 1979. Proto-Polynesian word list II. *Working Papers in Anthropology, Archaeology, Linguistics, and Maori Studies,* No. 53. Auckland: Department of Anthropology, University of Auckland.

Biggs, Bruce, D. S. Walsh, and Jocelyn Waqa. 1970. Proto-Polynesian reconstructions with English to Proto-Polynesian finder list, Interim Listing January 1970. Department of Anthropology, University of Auckland.

Blust, Robert, and Stephen Trussel. In progress. Austronesian comparative dictionary. Online at http://www.trussel2.com/acd.

Churchward, C. Maxwell. 1959. *Tongan Dictionary*. London: Oxford University Press.

Dryer, Matthew S., and Martin Haspelmath, eds. 2011. The world atlas of language structures online. Munich: Max Planck Digital Library.

Elbert, S. H. 1953. Internal relationships of the Polynesian languages and dialects. *Southwestern Journal of Anthropology* 9:147–73.

Ellis, L. B. M., and T. K. Attwood. 2001. Molecular biology databases: Today and tomorrow. *Drug Discovery Today* 6:509–13.

Fairbairn, D. J. 2011. The advent of mandatory data archiving. *Evolution* 65:1–2.

Forster, Johann Reinhold. 1778. A comparative table of the various languages in the isles of the South-Sea. In *Observations made during a voyage round the world*. London: G. Robinson. [ed. Nicholas Thomas et al., University of Hawai‘i Press, 1996, 188–89.]

Greenhill, Simon J., Robert Blust, and Robert D. Gray. 2008a. The Austronesian basic vocabulary database. Online at http://language.psy.auckland.ac.nz.

———. 2008b. The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics* 4:271–83.

Hale, Horatio. 1846. Essay at a lexicon of the Polynesian language. In *United States Exploring Expedition: Ethnography and Philology*, by Horatio Hale, 291–339. Philadelphia: Lea & Blanchard. [Reprinted Ridgewood, NJ: Gregg Press, 1968.]

Haspelmath, Martin, and Uri Tadmor. 2009. World loanword database. Online at http://wold.livingsources.org.

Kirch, Patrick Vinton, and Roger C. Green. 2001. *Hawaiki, ancestral Polynesia: an essay in historical anthropology*. Cambridge: Cambridge University Press.

Lemaître, Yves. 1973. *Lexique du tahitien contemporain*. Paris: ORSTOM.

Lynch, John. 2001. Too much to swallow: On terms meaning ‘swallow’ in Oceanic languages. *Oceanic Linguistics* 40:336–41.

———. 2002a. Proto-Oceanic labiovelars: Some new observations. *Oceanic Linguistics* 41:310–62.
———. 2002b. Potent roots and the origin of *kava*. *Oceanic Linguistics* 41:493–513.
Marck, Jeff. 1996. Was there an early Polynesian "Sky Father"? *Journal of Pacific History* 31:8–26.
———. 2000. *Topics in Polynesian language and culture history.* Canberra: Pacific Linguistics.
Nettle, Daniel, and Suzanne Romaine. 2000. *Vanishing voices: The extinction of the world's languages*. Oxford: Oxford University Press.
Pawley, Andrew. 2001. Bruce Biggs, 1921–2000: A tribute. *Oceanic Linguistics* 40:1–19.
Piwowar, H. A., R. S. Day, and D. B. Fridsma. 2007. Sharing detailed research data is associated with increased citation rate. *PLoS ONE* 2, e308. Online at www.plosone.org.
Pratt, George. 1911. *Pratt's grammar & dictionary of the Samoan language*. 4th ed. Apia, Western Samoa: Malua Printing Press.
Rausher, M. D., M. A. McPeek, A. J. Moore, L. Riseberg, and M. C. Whitlock. 2010. Data archiving. *Evolution* 64:603–4.
Relandus, Hadrianus. 1708. Dissertatio de linguis insularum quarundam orientalium. In *Dissertationum miscellanearum pars tertia,* by Hadrianus Renaldus, 57–139. Utrecht: G. Broedelet.
Rensch, Karl H. 2000. *The language of the Noble Savage*. Canberra: Archipelago Press.
Ross, Malcolm. 1995. Proto-Oceanic terms for meteorological phenomena. *Oceanic Linguistics* 34:261–304.
Rutter, Lawrence Kenji. 2001. A reevaluation of Proto-Polynesian *h. *Oceanic Linguistics* 40:20–32.
Soares, Pedro, Teresa Rito, Jean Trejaut, Maru Mormina, Catherine Hill, Emma Tinkler-Hundal, Michelle Braid, Douglas J. Clarke, Jun-Hun Loo, and Noel Thomson. 2011. Ancient voyaging and Polynesian origins. *American Journal of Human Genetics* 88(2):239–47.
Stimson, J. Frank, and Donald Stanley Marshall. 1964. *A dictionary of some Tuamotuan dialects of the Polynesian language*. The Hague: Martinus Nijhoff.
Tregear, Edward. 1891. *The Maori-Polynesian comparative dictionary.* Wellington: Lyon & Blair. (http://www.nzetc.org/tm/scholarly/tei-TreMaor.html).
Walsh, D. S., and Bruce Biggs. 1966. *Proto-Polynesian word list I*. Te Reo Monograph. Auckland: Linguistic Society of New Zealand.
Williams, H. W. 1971. *A dictionary of the Maori language.* Wellington: Government Printer.

s.greenhill@auckland.ac.nz r.clark@auckland.ac.nz